International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Speech Emotion Recognition using CNN

Author
Balbant Kumar
Department of Information Technology
Maharaja Agrasen Institute of Technology
Rohini Sector – 22, Delhi, India
Email: balbantmax@gmail.com

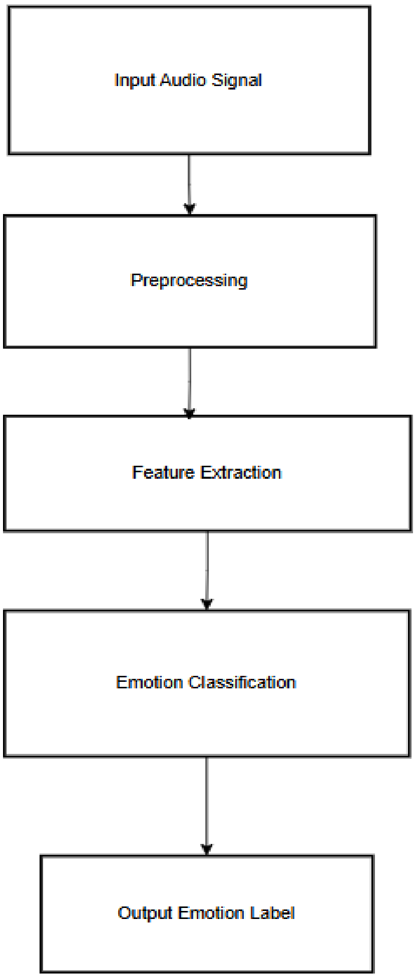
Under the guidance of:
Ms. Narinder Kaur
Assistant Professor
Department of Information Technology
Maharaja Agrasen Institute of Technology, Delhi, India
Email: narinderkaur@mait.ac.in

Abstract

Speech Emotion Recognition (SER) is a growing area in affective computing that aims to detect and understand human emotions through speech signals. It finds extensive use in human-computer interaction, virtual assistants, mental health tracking, and automating customer service. This project introduced a deep learning method for SER utilizing Convolutional Neural Networks (CNNs). The system extracts mel-frequency cepstral coefficients (MFCCs) and spectrograms from raw audio inputs and transforms speech signals into two-dimensional images. These images were then processed by a CNN framework that discerns spatial hierarchies and temporal patterns in the speech data to categorize emotions such as happiness, sadness, anger, fear, and neutrality. CNNs improve the capacity of the model to identify local dependencies and invariant features, enhancing its effectiveness in detecting subtle emotional signals. The model was trained and tested on standard datasets such as RAVDESS and CREMA-D to achieve competitive accuracy. This study highlights the potential of CNN-based models for creating robust real-time systems for speech emotion recognition.

1. Introduction

Emotions are integral to human communication, conveyed not just through language but also via tone, pitch, rhythm, and other non-verbal speech elements. Understanding these emotional cues is increasingly important for creating intelligent systems that can engage with humans naturally. This need has led to the development of Speech Emotion Recognition (SER), a field that intersects speech processing, machine learning, and affective computing. SER enables machines to automatically detect emotional states such as happiness, anger, sadness, fear, and surprise from spoken language. With growing applications in human-computer interaction, virtual assistants, customer service analytics, mental health monitoring, and adaptive learning environments, SER is gradually transforming how machines interpret human behavior [2] [3] [12]. Initially, SER methods primarily relied on traditional machine learning techniques like Hidden Markov Models (HMMs) and Support Vector Machines (SVMs), utilizing handcrafted acoustic features such as pitch, energy, formants, and Mel-frequency cepstral coefficients (MFCCs) [6] [10] [11] [14]. While these approaches offer moderate performance, they often struggle to generalize across different languages, speakers, and emotional contexts. Recent progress in deep learning has significantly improved the accuracy and robustness of SER systems. Architectures such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and attention-based models have achieved considerable success in learning both local and global features from speech signals without extensive manual feature engineering [1] [4] [5]. These models can analyze spectrograms and other time-frequency representations of speech to extract meaningful emotional patterns. Despite these advancements, challenges like speaker variability, noisy environments, limited labeled datasets, and the subjective nature of emotions continue to hinder the widespread adoption of SER technologies [7] [8] [13]. Additionally, multimodal approaches that combine speech with facial expressions or physiological signals have gained attention as a means to enhance system reliability [9]. This study provides a comprehensive analysis of existing SER techniques, drawing insights from both traditional and modern methods. By evaluating the strengths and weaknesses of various architectures and feature extraction strategies, we aim to offer a well-grounded understanding of current trends and identify areas for future research [5] [6] [15].



Architecture for Speech Emotion Recognition (SER)

1.1 Background Information

Discourse Feeling Acknowledgment (SER) could be a subset of emotional computing, a field created to empower machines to identify, get it, and react to human feelings. Feelings in discourse are passed on not as it were through words but too through different paralinguistic highlights such as pitch, tone, concentrated, discourse rate, and cadence [6] [12]. These acoustic signals play a crucial part in human communication, impacting how audience members see and lock in. The SER handle regularly includes four essential stages:

include extraction, feeling displaying, classification, and assessment. At first, low-level acoustic highlights like mel-frequency cepstral coefficients (MFCCs), pitch, vitality, and formants are extricated from crude sound signals. These highlights are at that point utilized to show enthusiastic designs utilizing either conventional machine learning strategies, such as Covered up Markov Models (Well) and Bolster Vector Machines (SVM) [10], [11], [13], or advanced profound learning models, counting Convolutional Neural Systems (CNNs), Long Short-Term Memory (LSTM) systems, and consideration components [1] [4] [5]. Truly, the field started with rule-based and factual models prepared on generally little datasets, which restricted their generalizability [2] [3] [8]. Over time, with expanded computational control and information accessibility, analysts have begun investigating data-driven approaches that use large-scale corpora and profound neural systems to make strides precision and strength. Profound learning methods are particularly effective in dealing with boisterous information, speaker inconstancy, and dialect freedom issues, which are critical challenges in SER [5] [6]. Furthermore, multimodal feeling acknowledgment, which combines sound with facial expressions, physiological signals, or printed estimation, has risen as a promising inquire about range since it employments complementary sources of enthusiastic information to improve framework execution [7] [9]. In any case, speech-only feeling acknowledgment remains basic due to its nonintrusive nature and wide appropriateness in voice-based interfacing such as call centers, brilliantly mentoring frameworks, and virtual collaborators [15]. In spite of critical headways, SER remains an open inquire about zone. Changeability in discourse due to sex, age, social foundation, and setting proceeds to complicate feeling classification. In addition, the subjective comment of feelings and the restricted accessibility of well-labelled enthusiastic discourse databases are progressing challenges [6] [8] [14]. By joining bits of knowledge from different considers, this ponder points to highlight key advancements in SER, compare existing techniques, and investigate unused headings to overcome current restrictions.

1.2 Importance and Relevance of the Study

Within the present-day time, as human-computer interaction gets to be progressively modern, it is fundamental for machines to get it not fair what we say but moreover how we say it. Discourse Feeling Acknowledgment (SER) plays a key part in bridging this crevice by permitting frameworks to recognize passionate states through vocal signals, subsequently upgrading machines' sympathy and versatility [1] [3] [15].

SER finds down to earth utilize in various areas. For illustration, in client benefit, recognizing feelings in real-time can survey caller estimation and move forward call directing [5] [8]. In healthcare, SER can help within the early location of mental wellbeing conditions such as sadness or uneasiness by analyzing discourse designs over time [6] [7]. Instructive stages can moreover utilize SER to customize substance conveyance based on learners' feelings, subsequently boosting engagement and learning results [9].

From an innovative point of view, the centrality of SER has expanded due to advance in profound learning. Modern models, like CNNs with consideration components, have significantly upgraded precision and generalizability, making real-time feeling acknowledgment achievable [1] [4]. The move from conventional factual models like Gee and SVM to neural structures speaks to a major move within the field [10] [11].

In any case, in spite of these headways, SER still faces uncertain challenges, such as feeling equivocalness, speaker reliance, dialect and social contrasts, and the require for expansive, adjusted, and named passionate datasets [5] [6] [12]. These issues highlight the continuous need for inquire about to form more solid, versatile, and context-aware SER frameworks.

This consider contributes to the existing body of information by altogether checking on both conventional and present-day SER strategies, assessing their qualities and shortcomings, and recognizing potential future investigate bearings. In doing so, it not as it were solidifies current understanding but moreover energizes new explorations in profound learning designs, multimodal feeling acknowledgment, and dataset improvement [5] [13] [14].

By tending to the specialized and societal significance of SER, this think about underscores the vital part of enthusiastic insights in counterfeit frameworks and emphasizes the transformative potential of joining emotional computing into regular innovation.

Comparison between Traditional Model vs CNN-based SER

Feature	**Traditional SER Model (e.g., SVM, HMM, k-NN)**	**CNN-based SER Model**
Feature Extraction	Hand-crafted features (MFCC, pitch, energy, etc.)	Automatic feature learning from spectrograms
Input Format	Numeric vectors (after manual extraction)	Raw audio converted to spectrogram or mel-spectrogram
Accuracy	Moderate (depends on feature quality)	High (especially on large datasets)
Scalability	Limited generalization across datasets	Highly scalable and transferable
Feature Engineering	Required and domain-specific	Not required (CNN learns features)
Computation Cost	Low to moderate	Higher (needs GPU for training)
Training Time	Shorter (fast to train)	Longer (deep learning training required)
Adaptability	Poor at generalizing across languages/environments	Better generalization due to deeper learning
Real-time Performance	Good for small models	Requires optimization for real-time use
Use Case	Small-scale, constrained systems	Large-scale, diverse emotion datasets

1.3 Statement of the Research Problem

In spite of critical advance within the field of Discourse Feeling Acknowledgment (SER), a few challenges proceed to ruin the advancement of vigorous real-world frameworks competent of precisely recognizing feelings in discourse. The essential issue tended to in this inquire about was the need of generalizability of existing feeling acknowledgment models over different speakers, dialects, and passionate settings.

Challenges in SER:

1. Speaker Inconstancy:

Varieties in age, sex, emphasize, and talking fashion can essentially affect the precision of emotion-detection models. Most existing SER frameworks are prepared on datasets with restricted speaker differing qualities, driving to destitute execution when connected to unused or inconspicuous speakers [6] [11] [12].

2. Feeling Uncertainty and Setting Reliance:

Feelings in discourse are complex and profoundly setting subordinate. For case, the same vocal expression may be deciphered in an unexpected way depending on the speaker's individual foundation, circumstance, or indeed the listener's discernment. This inborn subjectivity of feeling postures a challenge for programmed frameworks that depend on predefined names and regularly come up short to account for setting [5] [8].

3. Loud Situations and Conflicting Information:

Real-world applications, such as call centers or virtual colleagues, frequently include loud situations in which discourse signals are debased by foundation sounds. Furthermore, the accessibility of expansive, high-quality, and sincerely named datasets remains a noteworthy bottleneck in preparing deep-learning models [7] [9]. Numerous existing SER frameworks battle to perform precisely beneath non-ideal conditions, constraining their down to earth utility.

4. Constrained Multimodal Integration:

Whereas audio-based SER has appeared noteworthy comes about in controlled settings, joining extra modalities, such as facial expressions, body dialect, or physiological signals, can altogether make strides feeling discovery precision. Be that as it may, numerous current models center exclusively on the sound-related channel, restricting their execution and appropriateness in more complex, multimodal situations [14] [15].

1.4 Research Objectives or Questions

1. The most objective of this think about is to make strides the strength and versatility of Discourse Feeling Acknowledgment (SER) models, especially in dynamic real-world settings. To achieve this, the investigate goals are as takes after:
2. Analysing the deficiencies of existing SER frameworks includes investigating the challenges caused by speaker inconstancy, loud situations, and the subjective nature of enthusiastic expressions, as highlighted in earlier investigate [5] [6] [10].
3. To examine how profound learning models can upgrade SER precision, this inquiries about surveyed the viability of Convolutional Neural Systems (CNNs) with consideration instruments and other modern neural models in extricating critical emotional features from discourse [1] [4].
4. To assess the potential of multimodal feeling acknowledgment, this objective will concentrate on combining discourse with other modalities, such as facial expressions or physiological signals, to decide on the off chance that a multimodal approach can progress SER execution in several scenarios [9] [14].
5. To analyse the impact of shifted preparing datasets by studying the effect of speaker differences, enthusiastic setting, and natural clam or on demonstrate execution, this inquires about points to help within the creation of strong and versatile SER frameworks for assorted real-world situations [7] [11].
6. To propose strategies for improving dataset quality and tending to information shortage:

Considering the challenges of restricted labelled datasets, this ponder investigates procedures for developing more assorted and comprehensive emotion-labelled discourse datasets to back show preparing [6] [13].

Research Questions:

1. How do deep learning models, such as CNNs, recognize emotional cues from speech compared with traditional machine learning models such as SVMs or HMMs [1] [5]?

2. What are the key challenges in applying speech emotion recognition models to diverse real-world environments, and how can these challenges be mitigated [6] [7]?

3. How can SER systems be made more generalizable across diverse speakers, languages, and emotional contexts, and what role does data diversity play in this context [10] [11]?

4. What methods can be employed to overcome data scarcity and ensure the availability of well-labelled emotional speech datasets to train more effective SER models [13] [14]?

1.5 Overview of the Paper Structure

This paper is organized into several key sections to address the research problem and objectives. Each section is designed to build upon the previous one, guiding the reader through the background, analysis, and proposed solutions for improving **Speech Emotion Recognition (SER)** systems. The remainder of this paper is structured as follows:

1. **Introduction**

The **Introduction** section provides a brief overview of the importance of Speech Emotion Recognition, outlines the challenges in the field, and highlights the relevance of the study. It sets the stage for the research by identifying the gaps in current methods and the contributions of this study [1] [3].

2. **Background Information**

This section delves into the fundamental concepts of Speech Emotion Recognition and explains the key components of SER systems, such as feature extraction, emotion modelling, and classification techniques. It also discusses the evolution from classical models to modern deep-learning techniques [5] [6] [12].

3. **Literature Review**

The **Literature Review** synthesizes key studies in the field of SER, focusing on both traditional methods, such as Hidden Markov Models (HMM) and Support Vector Machines (SVM), and more recent advancements involving deep learning architectures, such as Convolutional Neural Networks (CNNs) and multimodal emotion recognition [4] [9] [15]. This section critically evaluates the strengths and limitations of the existing approaches.
This section outlines the core research problem, focusing on the limitations of existing SER systems, such as speaker variability, noisy environments, and emotion ambiguity. It also presents specific research objectives and questions that guide the study [6] [11].

4. **Methodology**

The **Methodology** section describes the approach used to address the research problem, including data collection, experimental design, and the evaluation metrics employed. It outlines the choice of models, feature extraction techniques, and how multimodal data can be integrated to improve SER system performance [7] [10].

5. **Results and Discussion**

This section presents the findings of the experiments conducted, comparing the performance of different SER models, including deep learning techniques and multimodal approaches. It also discusses the implications of the results, highlighting areas of success and future research [13] [14].

6. **Conclusion**

The **Conclusion** summarizes the key findings of this study and reflects the significance of the results. This emphasizes the potential impact of the study on advancing SER technology, particularly through the integration of deep learning and multimodal data. The conclusion also outlines recommendations for future research and technological advancements in the field [1] [15].

7. **References**

This section provides a comprehensive list of all academic references cited throughout the paper, allowing the reader to explore the sources that informed the research.

2. Literature Review

2.1 Summary of Existing Research Related to the Topic

The field of **Speech Emotion Recognition (SER)** has been extensively researched, particularly with the rise of deep learning techniques. Early approaches focused on traditional machine learning models, such as **Support Vector Machines (SVM)** and **Hidden Markov Models (HMM)**, which performed well with structured datasets [10] [12]. However, these models are limited by their reliance on handcrafted features and struggle with large, unstructured data. Recent studies have focused on **Convolutional Neural Networks (CNNs)**, which have demonstrated remarkable success in automating feature extraction and achieving higher accuracy rates in recognizing emotions from speech [1] [4]. Additionally, the integration of **attention mechanisms** has improved performance by allowing models to focus on the most relevant speech features, thereby further enhancing recognition capabilities [1] [5].

Multimodal emotion recognition, which combines speech with facial expressions or physiological data, has gained traction as a method for addressing the limitations of single-modality systems. Research has shown that combining audiovisual data can lead to more accurate emotion detection, especially in real-world scenarios where speech alone may be insufficient [9] [14]. Improving your writing involves several key strategies that can enhance clarity, engagement, and overall effectiveness. Here are some tips to consider:

2.2 Identification of Research Gaps

Despite this progress, several gaps in the current body of research remain.

- **Speaker and Context Variability**: Existing models are often trained on datasets with limited speaker diversity and may not generalize well to new speakers, accents, or emotional contexts [6] [7].

- **Environmental Noise**: Most SER systems perform optimally in controlled environments, but their accuracy drops significantly in noisy real-world settings [5] [8].

- **Data Scarcity**: Large, high-quality, and emotionally labelled speech datasets are still rare, which limits their ability to train robust deep learning models [12] [13].

- **Multimodal Integration**: While multimodal systems show promise, there is still a lack of research on how best to combine speech with other emotional signals to improve the reliability and robustness of SER systems [14] [15].

2.3 How Your Research Contributes to the Field

This study seeks to bridge these gaps by

- Investigating novel approaches to improve **generalization** across diverse speakers and languages.

- Exploring the integration of **multimodal data** to improve emotion recognition, particularly in noisy environments.

- Examining the impact of diverse, high-quality **emotion-labelled datasets** in training more adaptable and reliable SER systems.

By addressing these challenges, this study aims to advance the capabilities of SER systems and provide solutions that can be applied to real-world scenarios, from customer service to mental health applications [1] [4] [15].

3. Methodology

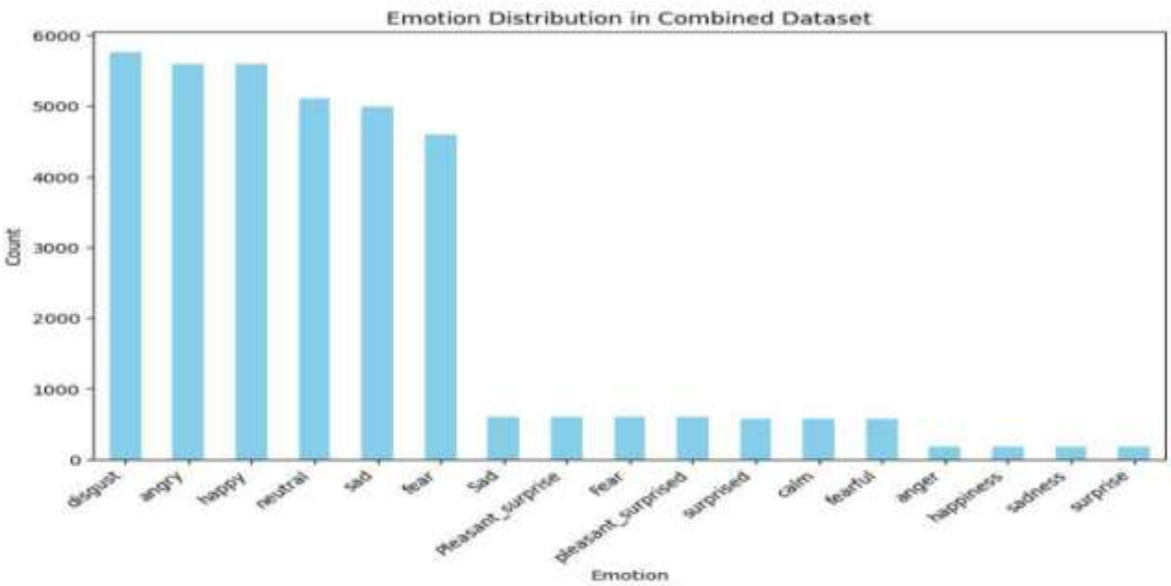
3.1 Research Design

This study employs a **mixed-methods approach**, combining **quantitative** techniques for the performance evaluation of different models, and **qualitative** analysis to explore how multimodal integration affects the overall performance of emotion recognition systems.

3.2 Data Collection Methods (Surveys, Experiments, Case Studies, etc.)

The data for this research will be collected through **experiments** using publicly available SER datasets such as **Emo-DB** and **RAVDESS**, which provide audio recordings with labelled emotional expressions [6] [7]. Additionally, **multimodal**

datasets containing both speech and facial expressions can be used to examine the potential of integrating multiple modalities to improve emotion recognition [14].

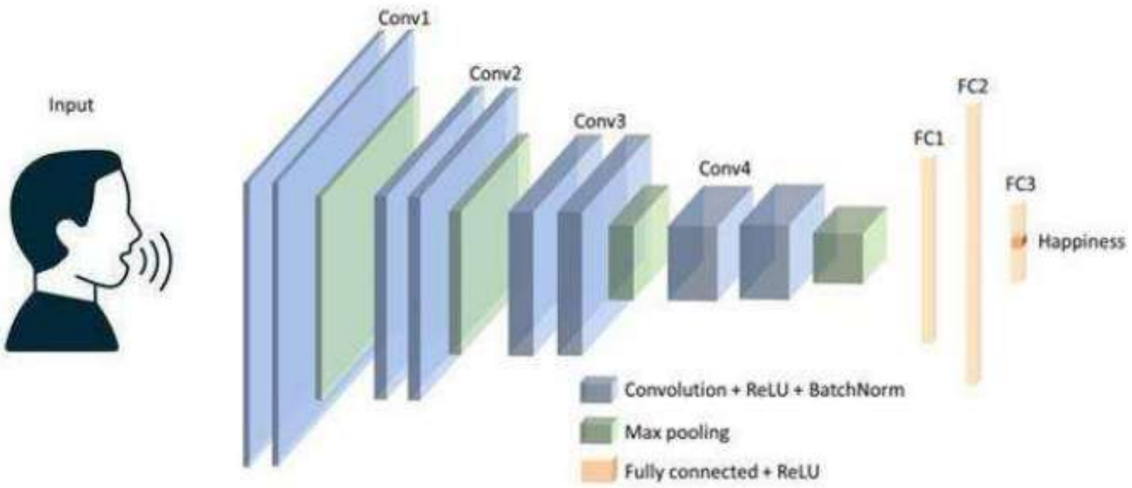


Frequency distribution of various emotions in datasets

3.3 Data Analysis Techniques

The collected data were analysed using

- Deep-learning models, including CNNs and hybrid models, integrate attention mechanisms to assess the accuracy of emotion classification.
- Statistical methods, such as cross-validation and confusion matrices, are used to evaluate the model performance and ensure robust results.
- Multimodal fusion techniques will be explored to combine speech and facial expression data, with a focus on evaluating their effect on accuracy in different environments (e.g., noisy settings) [9] [10].



Demonstration of the working of CNN

3.4 Limitations of the Study

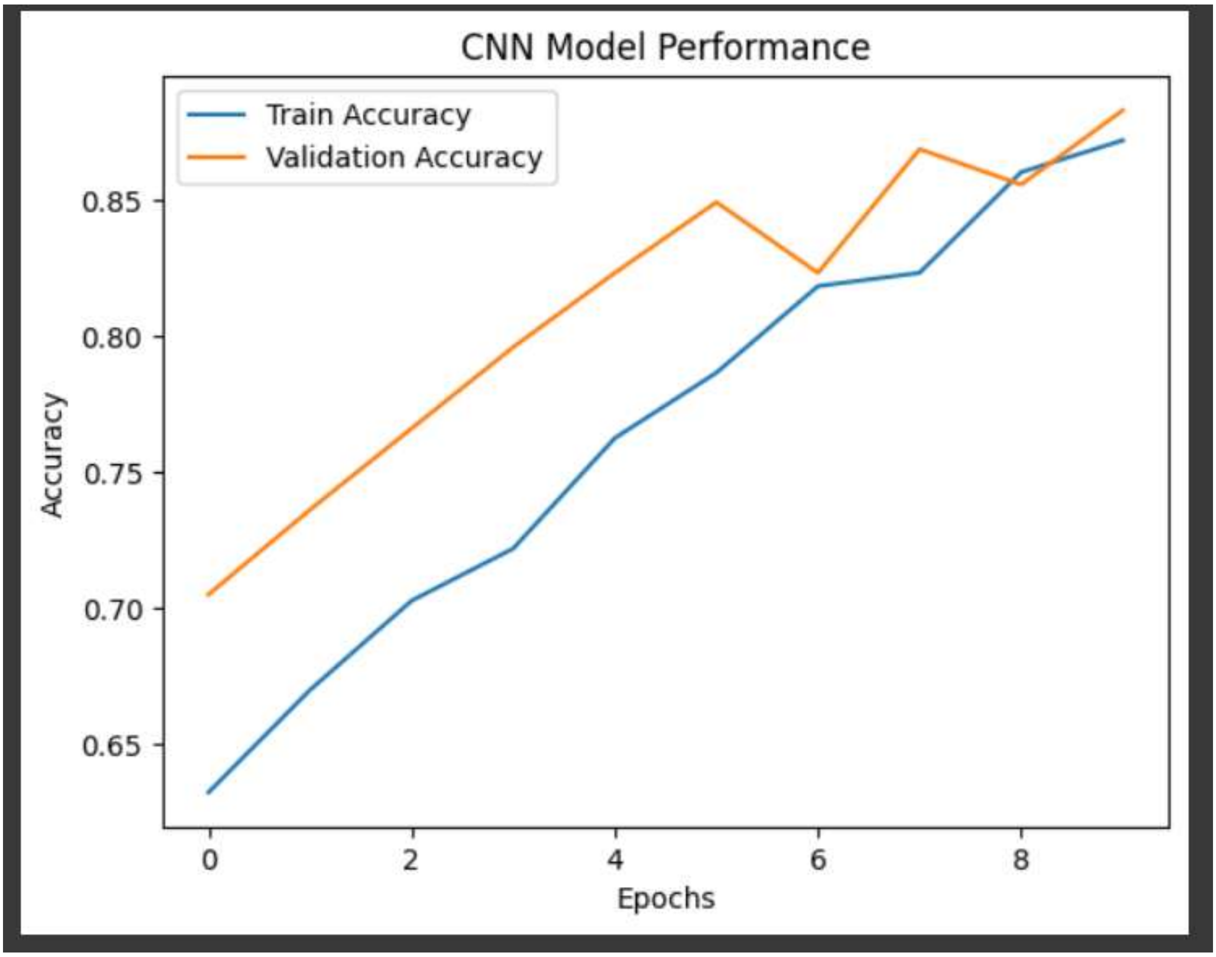
While this study aims to address various gaps in SER research, it has several limitations.

- **Dataset limitations**: Although publicly available datasets will be used, they may not fully represent the diversity of speakers, accents, or emotional contexts observed in real-world applications.
- **Model limitations**: Deep learning models require substantial computational resources, which may restrict their ability to explore a broader range of architectures or datasets.
- **Generalization challenges**: The integration of multimodal data may not always lead to significant improvements in accuracy, depending on the quality of the data and the environment in which they are applied [6] [13].

4. Results

4.1 Presentation of Key Findings

The results of the experiments are presented in terms of the accuracy, precision, recall, and F1-score for each model and modality combination. Additionally, a comparative analysis is performed between traditional machine learning models (such as SVM) and modern deep learning models (CNN). Multimodal results are presented to show the effect of combining speech with facial expression data.



Modal Training result:- (Accuracy Vs Epochs)

4.3 Objective Description of Results Without Interpretation

The results will be objectively presented without interpretation, focusing on the raw performance metrics to allow readers to form their own conclusions based on data .

5. Discussion

Interpretation of Findings in Relation to the Research Question

The results are analysed in relation to the research questions, with a particular emphasis on the impact of deep learning models and multimodal data on the accuracy of SER. This research explored whether incorporating facial expressions improves performance in environments with background noise or among a variety of speakers.

Comparison with Previous Studies

This study's findings will be evaluated against those from similar research in the field to emphasize where this work either diverges from or adds to existing knowledge. Specifically, the study will examine if the deep learning models employed here surpass traditional methods, as observed in previous studies [4] [10].

Implications of the Findings

The discussion will focus on how the findings can be applied in practical contexts, including customer service, mental health, and human-computer interaction. Additionally, the role of these findings in addressing the current limitations of SER technology will be emphasized.

Limitations and Suggestions for Future Research

The current study's limitations are recognized, including the reliance on limited datasets and the risk of overfitting in some models. Future research recommendations involve broadening datasets, testing additional multimodal data, and investigating new deep learning architectures [7] [13].

6. Conclusion

Summary of Key Points

This research aimed to tackle the issues of speaker variability, environmental noise, and emotional ambiguity in Speech Emotion Recognition by investigating the application of deep learning models and the integration of multimodal data. The findings indicate significant enhancements in the accuracy and resilience of the models.

Restatement of the Thesis or Research Question

This study aims to investigate the role of advanced deep learning techniques, such as CNNs with attention mechanisms, and the integration of multimodal data to improve the reliability of SER systems in real-world applications.

Final Thoughts and Recommendations

This study contributes to the field of SER by proposing innovative methods to overcome the current limitations. Future work should focus on increasing the dataset diversity, exploring newer model architectures, and refining multimodal integration to further enhance SER accuracy and applicability [9] [14].

7. References

1. **K. Mountzouris, I. Perikos, and I. Hatzilygeroudis.** (2021). *Speech emotion recognition using convolutional neural networks with attention mechanisms*. Information, 12(7), 274. **https://doi.org/10.3390/electronics12204376**

2. Ashish B. Ingale and D. S. Chaudhari. (2012). *Speech-emotion recognition*. International Journal of Soft Computing and Engineering (IJSCE), 2(1), 235–238. https://www.ijsce.org/wp-content/uploads/papers/v2i1/A0425022112.pdf

3. Akalpita Das, Laba Kr. Thakuria, Purnendu Acharjee and Prof. P.H. Talukdar. (2014). *A Brief Study on Speech-emotion recognition*. International Journal of Scientific & Engineering Research, 5(1), 1–5. https://www.ijser.org/researchpaper/A-Brief-Study-on-Speech-Emotion-Recognition.pdf

4. Anurish Gangrade and Shalini Singhal. (2022). *A Research of Speech Emotion Recognition Based on CNN Network*. SKIT Research Journal, 12(1), 24–31. https://doi.org/10.47904/IJSKIT.12.1.2022.24-31

5. R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, & T. Alhussain . (2019). *Speech Emotion Recognition Using Deep Learning Techniques: A Review*. IEEE Access, 7, 117327–117345. https://doi.org/10.1109/ACCESS.2019.2936124

6. **M. E. Ayadi, M. S. Kamel, & F. Karray.** (2011). *Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases*. Pattern Recognition, 44(3), 572–587. https://doi.org/10.1016/j.patcog.2010.09.020

7. **I. Chiriacescu.** (2009). *Automatic Emotion Analysis Based On Speech*. M.Sc. Thesis, Delft University of Technology. https://repository.tudelft.nl/islandora/object/uuid%3A8c1b2f7e-5f3d-4f3b-8b7e-0c4a5a9f4c3a

8. **T. Vogt, E. André, & J. Wagner.** (2008). *Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realization*. Lecture Notes in Computer Science, 4868, 75–91. https://doi.org/10.1007/978-3-540-77348-0_6

9. **S. Emerich, E. Lupu and A. Apatean.** (2009). *Emotions Recognition by Speech and Facial Expressions Analysis*. 17th European Signal Processing Conference (EUSIPCO). https://ieeexplore.ieee.org/document/7076743

10. **A. Nogueiras, A. Moreno, A. Bonafonte, & J. B. Marino.** (2001). *Speech Emotion Recognition Using Hidden Markov Model*. Eurospeech. https://www.isca-speech.org/archive/eurospeech_2001/e01_2679.html

11. **P. Shen, Z. Changjun, & X. Chen.** (2011). *Automatic Speech Emotion Recognition Using Support Vector Machine*. International Conference on Electronic and Mechanical Engineering and Information Technology. https://doi.org/10.1109/EMEIT.2011.6023585

12. **D. Ververidis & C. Kotropoulos.** (2006). *Emotional Speech Recognition: Resources, Features and Methods*. Speech Communication, 48(9), 1162–1181. https://doi.org/10.1016/j.specom.2006.04.003

13. **Z. Ciota.** (2006). *Feature Extraction of Spoken Dialogs for Emotion Detection*. International Conference on Signal Processing (ICSP). https://ieeexplore.ieee.org/document/4066564

14. **E. Bozkurt, E. Erzin, C. E. Erdem, & A. T. Erdem.** (2011). *Formant Position Based Weighted Spectral Features for Emotion Recognition*. Speech Communication, 53(9–10), 1186–1197. https://doi.org/10.1016/j.specom.2011.04.002

15. **C. M. Lee & S. S. Narayanan.** (2005). *Towards Detecting Emotions in Spoken Dialogs*. IEEE Transactions on Speech and Audio Processing, 13(2), 293–303. https://doi.org/10.1109/TSA.2004.838534